PROTAX-SSL: PROBABILISTIC TAXONOMIC CLASSIFICATION WITH SELF-SUPERVISED REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Taxonomic classification of DNA barcodes is fundamental to biodiversity monitoring, yet remains challenging due to class imbalance, incomplete databases, and unknown taxa. In such settings, reliable taxonomic inference requires calibrated uncertainty estimates and identifying unknown taxa. Self-supervised foundation models have demonstrated high-order representations that capture taxonomic relationships, but downstream classifiers built on them typically lack principled uncertainty estimates and mechanisms for modeling unknown taxa. PROTAX, a probabilistic classifier from statistical ecology, provides calibrated predictions and explicitly models taxonomic structure complexities, but operates on raw sequences, limiting its accuracy relative to self-supervised approaches. We bridge these two lines of work with PROTAX-SSL, which couples self-supervised DNA barcode representations with probabilistic taxonomic inference. We evaluate variants of PROTAX-SSL across three datasets spanning diverse experimental settings. PROTAX-SSL improves calibration and unknown-species detection, while its modular design leverages complementary representations to improve accuracy. We provide a basis for reliable taxonomic analysis and potential species discovery.

1 INTRODUCTION

Biodiversity sustains ecosystem stability and human well-being, yet it is declining under anthropogenic pressure, leading to widespread species loss and ecosystem degradation (Pimm et al., 2014; Sala et al., 2000). Accurately estimating global biodiversity remains a fundamental challenge, as traditional taxonomic analysis methods are time-consuming, rely on expert knowledge, and do not scale effectively to the large collection and analysis of specimens required for comprehensive biodiversity monitoring (Gautam et al., 2022). Moreover, a substantial proportion of Earth's biodiversity remains unknown, with an estimated 86% of terrestrial and 91% of marine species not formally described (Mora et al., 2011). This presents a challenge for DNA-based biodiversity assessment, where classifiers typically assign DNA sequences to known taxonomic groups inside reference databases, and therefore assume that each query belongs to a known taxon (Ross et al., 2008; Ye et al., 2019). Since collected samples may originate from unknown species or species lacking representative sequences (Keck et al., 2023), detection of such novel taxa becomes essential for accelerating species discovery and biodiversity understanding (Bickford et al., 2007; Pimm et al., 2014). Reliable predictive probabilities are also important in this setting, as classification accuracy alone provides no indication of how much confidence should be placed in an individual assignment. With incomplete reference databases and continually evolving taxonomic knowledge, incorrect predictions are unavoidable and unreliable probabilities may undermine ecological policy decisions (Singh et al., 2024).

DNA barcodes are short, standardized genetic sequences for specimen identification (Hebert et al., 2003), with Cytochrome c oxidase Subunit I (COI) and Internal Transcribed Spacer (ITS) markers for animals and fungi, respectively (Hollingsworth et al., 2009; Schoch et al., 2012). This approach scales where morphological identification does not (Dopheide et al., 2019), but the underlying data remain difficult. Arthropods and fungi are among the most diverse and taxonomically complex groups of life, characterized by highly imbalanced distributions and substantial gaps in species annotations (Basset et al., 2012; Romeijn et al., 2024). Fungi are more challenging due to limited diagnostic morphological features, forcing near-complete reliance on DNA (Bickford et al., 2007).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Traditional classification tools, such as BLAST (Altschul et al., 1990), the RDP classifier (Wang et al., 2007), and `dnabarcoder` (Vu et al., 2022), perform well when taxa are sufficiently represented in the reference database, but degrade under conditions of incomplete and imbalanced data (Austerlitz et al., 2009; Bazinet & Cummings, 2012). They also exhibit prohibitive inference times at scale and limited generalization to unseen taxa (Ross et al., 2008; Ye et al., 2019). More recently, deep learning approaches such as convolutional neural networks (CNNs) and transformer-based architectures demonstrated strong performance in fully supervised settings (Badirli et al., 2021; Romeijn et al., 2024), but rely densely on labeled data and are therefore sensitive to the class imbalance and label sparsity characteristic of arthropod and fungal datasets.

Foundation models have shown that self-supervised learning (SSL) yields transferable representations across downstream tasks by leveraging large datasets without annotations, making them well-suited to biodiversity. Building on these advances, DNA barcode foundation models such as BarcodeBERT (Millan Arias et al., 2026) and BarcodeMamba (Gao & Taylor, 2024) have been developed for taxonomic classification and demonstrated strong performance, in many cases surpassing existing state-of-the-art methods. The embedding spaces learned by these models capture taxonomic structures and exhibit strong correspondence with known taxonomic relationships. However, limitations remain in how these representations are exploited for taxonomic classification. Downstream classifiers typically rely on linear probes or 1-nearest-neighbor approaches (Millan Arias et al., 2026; Gao & Taylor, 2024), which do not model taxonomic hierarchy, and assume queries belong to known reference taxa, thereby limiting unknown taxa identification. More importantly, they lack principled uncertainty estimates: cosine similarities or confidence scores treated as probabilities yield overconfident errors that, combined with incomplete databases and within-species variability (Somervuo et al., 2017), lead to unreliable classifications. Studies have shown that modern deep learning models can achieve high accuracy while remaining biased in their predictions (Guo et al., 2017).

Developed in statistical ecology rather than deep learning, PROTAX (Somervuo et al., 2016) is a well-established probabilistic framework specifically designed for taxonomic inference. Unlike conventional classifiers, it provides unbiased probabilities of specimen identification and explicitly accounts for incomplete reference libraries and unknown taxa. It also produces valid taxonomic paths consistent with the taxonomic hierarchy and can integrate heterogeneous sources of information within a unified framework. However, PROTAX has scalability limitations, which PROTAX-GPU (Li et al., 2024) addresses through a GPU-accelerated implementation of the same model. Nevertheless, both approaches rely on raw DNA barcode sequences and lack the ability to learn representations, limiting their ability to capture higher-order features relevant to taxonomic inference.

We summarize our main contributions as: (1) We propose PROTAX-SSL, a novel hybrid taxonomic classification framework that combines the expressive power of self-supervised representations with coherent taxonomic predictions, calibrated uncertainty estimates, and explicit modeling of unknown taxa. (2) We comprehensively evaluate DNA barcode foundation models, PROTAX-GPU, and PROTAX-SSL across three benchmark datasets spanning diverse experimental settings. We assess both classification accuracy and uncertainty calibration quality to provide a holistic evaluation of performance reliability. (3) We further evaluate our framework against PROTAX-GPU on the detection and hierarchical placement of unknown species in large reference libraries, assessing its ability to identify these sequences and correctly associate them with their closest taxonomic context.

2 PRELIMINARIES

In this section, we provide a brief explanation of the models upon which we build our approach.

2.1 PROTAX

PROTAX (Somervuo et al., 2016) is designed as a probabilistic taxonomic layer operating over heterogeneous sources of evidence. Rather than relying on a specific feature representation, it flexibly combines information from existing predictors, such as sequence similarity and outputs from other classifiers, to produce calibrated taxonomic assignment probabilities. We briefly describe key concepts of PROTAX relevant to our framework below.

PROTAX operates on a reference taxonomy represented as a hierarchical tree, where taxonomic units at different levels are represented as nodes and parent–child relationships define the struc-

ture. Reference sequences are associated with the finest-grained available taxonomic unit to which they are assigned, with each node potentially having zero, one, or multiple sequences. At each node, PROTAX adds an "unknown" branch alongside the known ones to represent the possibility of queries belonging to an unknown taxon under that node. Prediction proceeds hierarchically from the root toward the leaves, considering only the children of the previously selected parent node at each step.

Through this hierarchical classification, PROTAX constructs a node-specific design matrix at each step. The design matrix encodes both taxonomic and sequence-derived information for each candidate branch. Two taxonomic categorical features describe the state of the branch within the reference taxonomy, such as whether it represents a known taxon and has reference sequences available, enabling the model to reason about the underlying taxonomy. Sequence-derived features capture relationships between the query and reference sequences, such as measures of sequence similarity.

A Bayesian multinomial regression maps the design matrix to a probability distribution over candidate branches. For node z with n_z children, the probability of assigning a query sequence to child i is

$$P(Y_z = i) = \frac{\exp\left(\sum_{j=1}^{m} X_{ij}^{z} \beta_j^{l}\right)}{\sum_{k=0}^{n_z} \exp\left(\sum_{j=1}^{m} X_{kj}^{z} \beta_j^{l}\right)}$$

where $X^z \in \mathbb{R}^{n_z \times m}$ is the design matrix at node z, with m features over n_z child branches. The only learned model parameters are the regression coefficients β^l, which are shared across taxonomic level l and learned independently for each level, allowing the model to capture level-specific differences in feature reliability. Learning the appropriate parameters yields probabilities that correspond to empirical frequencies of correct classification, providing an interpretable measure of uncertainty.

PROTAX-GPU. (Li et al., 2024) proposed an extension of PROTAX designed for large reference databases by replacing the Bayesian procedure with gradient-based optimization using cross-entropy loss. To obtain sequence-derived features, it computes the fractional dissimilarity (p-distance) (Nei & Kumar, 2000) between the query and reference sequences, from which two features are extracted to populate the design matrix: (1) the minimum distance between the query and all reference sequences assigned to a taxon, and (2) the gap between the minimum and second-minimum distances.

2.2 DNA Barcode Encoders

Once foundation models are pretrained, they can be used as encoders to transform input sequences into high-order representations, which can then be adapted to downstream tasks. In this work, we use two state-of-the-art foundation models as DNA barcode encoders: BarcodeBERT (Millan Arias et al., 2026), a Transformer-based model trained using masked language modeling, and BarcodeMamba (Gao & Taylor, 2024), a state-space model (Dao & Gu, 2024) trained using next-token prediction. Using two architectures with fundamentally different inductive biases is intended to demonstrate the generalizability of our proposed framework across representation backbones.

3 Methodology

Our proposed framework builds upon PROTAX-GPU (Li et al., 2024) by replacing sequence dissimilarities computed from raw DNA barcodes with distances measured between self-supervised DNA barcode representations. The framework consists of two main pipelines. The first constructs the reference taxonomy by building the taxonomic tree, encoding reference barcode sequences, and associating the resulting embeddings with their corresponding taxonomic nodes, thereby establishing the reference base used for classification. The second pipeline classifies query sequences by encoding them into the same embedding space and comparing their embeddings with the reference embeddings to derive similarity-based features. These features are incorporated into the PROTAX design matrix and processed through the hierarchical probabilistic procedure described previously. Figure 1 provides a high-level overview of the PROTAX-SSL framework.

3.1 Barcode Embedding and Sequence-derived Features Extraction

Reference DNA barcode sequences are encoded using one or both of the pretrained barcode encoders considered in this work. Given a barcode sequence b, an encoder produces contextualized

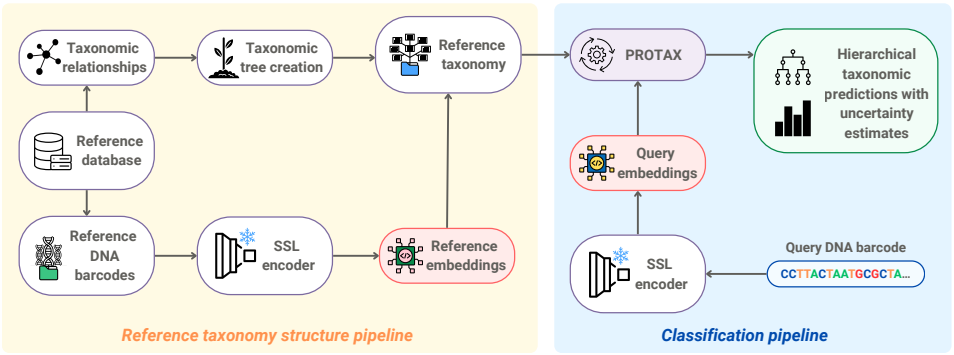

Figure 1: Overview of PROTAX-SSL, consisting of two pipelines: (left) reference taxonomy construction, and (right) taxonomic classification. Best viewed in color.

token representations $H(b) \in \mathbb{R}^{T \times d}$, which are aggregated by global average pooling to obtain a fixed-length embedding $e(b) = \frac{1}{T} \sum_{i=1}^{T} h_i$, where T is the number of tokens and $d = 768$ for both foundation models. Throughout all experiments, the pretrained barcode encoders are kept frozen and are not further fine-tuned on any downstream datasets. This design isolates the effect of the proposed framework and ensures that the learned representations reflect only the information acquired during pretraining. Consequently, PROTAX-SSL operates as a post-hoc probabilistic layer on top of existing DNA barcode foundation models rather than modifying the underlying sequence encoder.

During classification, given a query embedding $e(q)$ and a reference embedding $e(r)$, distances between query and reference embeddings are computed from cosine similarity:

$$\mathrm{sim}(q, r) = \frac{e(q) \cdot e(r)}{|e(q)||e(r)|}, \qquad d(q, r) = \frac{1}{2}\left(1 - \mathrm{sim}(q, r)\right).$$

This produces distances in $[0, 1]$. Using these embedding-based distances, the same two features are extracted for each candidate taxon as PROTAX-GPU (minimum distance and gap). These features are then concatenated with the taxonomic categorical features to form the input design matrices.

3.2 PROTAX-SSL Variants

The flexibility of PROTAX-SSL allows the framework to be instantiated with different combinations of foundation models. Where unambiguous, we omit "-SSL" from variant names for conciseness.

Linear Models. We consider three linear variants based on the choice of encoder used to generate sequence representations. `PROTAX-SSL[BarcodeBERT]` and `PROTAX-SSL[BarcodeMamba]` use a single embedding space, while `PROTAX-SSL[BarcodeBERT+Mamba]` combines both, computing embeddings and distances independently in each space before concatenating the extracted features to form the design matrix. As a result, the design matrix expands from $(N, 4)$ to $(N, 6)$, with corresponding regression parameters of shape $(l, 6)$ instead of $(l, 4)$. This formulation allows us to explore whether integrating information from both encoders yields improvement in predictive performance and calibration.

Multilayer Perceptron Model. We further introduce a fourth variant, `PROTAX[BarcodeBERT+Mamba, MLP]`, which replaces the linear regression component $\sum_{j=1}^{m} X_{ij}^{z} \beta_j^{l}$ with a multilayer perceptron (MLP), allowing the model to create more complex decision boundaries that better exploit features derived from multiple embedding spaces. Specifically, a separate MLP is trained for each taxonomic level, with all networks sharing the same architecture and hyperparameter configuration. Details of the MLP architecture are described in Appendix C.

For a taxonomic level l, let $X^z \in \mathbb{R}^{n_z \times 6}$ denote the node-specific design matrix, where X_i^z is the feature vector for branch i. The hidden representations of the (M) fully connected layers are:

$$h_{i1}^{z} = \phi\left(W_1^{l} X_i^{z} + b_1^{l}\right), \qquad h_{ik}^{z} = \phi\left(W_k^{l} h_{i(k-1)}^{z} + b_k^{l}\right), \quad k = 2, \ldots, M,$$

where $\phi(\cdot)$ denotes the ReLU nonlinearity. The final representation is mapped to a scalar $a_i^z = v^l h_{iM}^z + c^l$, and the resulting score is normalized over children using a softmax to obtain $P(Y_z = i)$.

4 Experimental Set-up

This section describes the datasets, evaluation tasks, and baselines used in our experiments. Training configurations and computational resources details are provided in Appendix B.

4.1 Datasets

We evaluate PROTAX-SSL across three datasets spanning different barcode markers, dataset scales, taxonomic completeness and cluster quality, to provide comprehensive evaluation settings. An overview of their taxonomic coverage is shown in Appendix A.

MycoAI. Introduced by (Romeijn et al., 2024), this large-scale fungal dataset contains approximately 5.23 million ITS barcodes from the UNITE+INSD database (Abarenkov et al., 2023), with severe class imbalance and 93% of sequences lacking species-level annotations. We use the MycoAI Benchmark, which contains the 7% of sequences fully identified to the species level, as our primary fungal training dataset, comprising approximately 367,000 sequences. We use the MycoAI validation split, a random 0.2% hold-out subset of the training data, for hyperparameter tuning.

In addition, we evaluate the models trained on the MycoAI Benchmark on two test sets: Yeast (Vu et al., 2016), containing approximately 4,400 ITS sequences, and Filamentous Fungi (Vu et al., 2019), containing approximately 11,600 sequences. To better characterize the difficulty of these evaluation settings, Appendix A summarizes the overlap between them and the training set.

To evaluate unknown species detection and taxonomic placement, we augment the training set with sequences from the full MycoAI dataset that are annotated only to the genus level, restricting them to genera represented in the MycoAI Benchmark. Following PROTAX (Somervuo et al., 2016), we randomly select a number of these sequences equal to 5% of the training set size and add them as representatives of unknown species within known genera. For each added sequence, a different genus-level sequence from the same genus is held out and used as an unknown species test example.

Canadian Invertebrates. Curated through the Barcode of Life Data System (Ratnasingham & Hebert, 2007), this large arthropod barcode library (deWaard et al., 2019) is provided in the preprocessed splits of Millan Arias et al. (2026), referred to as `CanadianInvertebrates-ML`. The dataset contains approximately 974,000 COI sequences exhibiting class imbalance and is divided into Pretrain, Seen, and Unseen subsets. Pretrain is used to pretrain BarcodeBERT and BarcodeMamba, while Seen contains approximately 67,000 fully identified sequences for supervised learning; we use its 70% training split for training PROTAX-SSL and its 20% test split for evaluation.

FinBOL. This highly curated arthropod barcode library (Roslin et al., 2022) contains approximately 37,000 expertly identified COI sequences. Its meticulous annotations and well-represented sequence space, characterized by low intra-species and high inter-species variability, make raw sequence-derived features highly informative for taxonomic classification. FinBOL therefore provides a complementary setting for evaluating whether the self-supervised representations used by PROTAX-SSL provide additional benefits over raw sequence similarity in such settings.

4.2 Evaluation Tasks

We evaluate model performance across three complementary dimensions, described below.

Taxonomic Classification Accuracy. This evaluation measures the correct assignment of a query sequence to its corresponding taxonomic units within the hierarchy. Accuracy is reported across three levels: family, genus, and species, with finer levels being more challenging. PROTAX-SSL constructs sequence-derived features by comparing queries against a reference library built from the training data. During training and evaluation, identical copies of the query are temporarily excluded

from the reference library to prevent trivial self-matches. For test sets with partial overlap, this self-match exclusion is applied to queries for which an identical copy exists. For results reported on the training data with self-match exclusion, this protocol evaluates generalization beyond observed sequences. PROTAX-SSL classification layers are trained on extracted low-dimensional sequence-derived features rather than the underlying sequence representations, while the barcode encoders, which account for most model parameters, are pretrained on external data and kept frozen, substantially limiting sequence-specific memorization. This protocol is consistent with the original PROTAX study (Somervuo et al., 2016), which trained and evaluated their model on the same dataset.

Predictive Reliability. We evaluate the calibration of predictive probabilities using cumulative calibration plots adopted by (Somervuo et al., 2016). We plot the cumulative sum of ascendingly sorted probabilities against the cumulative sum of correct predictions without binning, providing a fine resolution. A perfectly calibrated model follows the diagonal, while deviations above or below this line indicate underconfidence and overconfidence, respectively. We also report the expected calibration error (ECE) across confidence bins as $\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{n} \left|\mathrm{acc}(S_b) - \mathrm{conf}(S_b)\right|$, where B is the number of bins, S_b the predictions in bin b, n the total number of predictions, $\mathrm{acc}(S_b)$ the empirical accuracy of predictions in bin b, and $\mathrm{conf}(S_b)$ their mean predicted confidence.

Unknown Species. Finally, we evaluate PROTAX-SSL on two tasks involving unknown species: (1) detection, measuring the accuracy of identifying queries that do not belong to any species represented in the reference taxonomy, and (2) placement, measuring the accuracy of assigning model-flagged "unknown" species to the correct parent genus within the hierarchy.

4.3 Baselines

We compare PROTAX-SSL variants against the base models on which they are built. First, we compare against PROTAX-GPU to evaluate replacing the raw sequence space with self-supervised representations. Second, we compare against the underlying barcode encoders with k-NN as a classifier, using cosine similarity as a proxy for predictive uncertainty. This comparison evaluates the effect of replacing conventional downstream classifiers with PROTAX and tests whether the model can effectively exploit multiple embedding spaces. All of these methods, including PROTAX-SSL, are reference-based; therefore, self-match exclusion is applied to each of them.

Because fungal classification on MycoAI serves as our most challenging benchmark, we extend this evaluation to several state-of-the-art non-reference-based classifiers. These encompass the non-learning algorithmic BLAST (Altschul et al., 1990), alongside supervised deep learning models: taxonomic CNN model (Badirli et al., 2021), fine-tuned BarcodeBERT (Millan Arias et al., 2026), MycoAI-CNN, and MycoAI-BERT (Romeijn et al., 2024). Results for these baselines are taken directly from Gao et al. (2025), whose evaluation setup closely mirrors ours. Evaluating against these distinct paradigms contextualizes the performance of PROTAX-SSL beyond reference-based approaches. However, we treat this as an indirect comparison, since the non-reference baselines in Gao et al. (2025) were evaluated without self-match exclusion and trained directly on MycoAI, whereas PROTAX-SSL relies on frozen encoders pretrained on arthropod COI data.

5 Results and Discussion

We split our experiments across arthropods and fungi. Since encoders were pretrained on arthropod barcodes, fungal classification acts as an out-of-domain benchmark, testing if PROTAX-SSL provides benefits when representations degrade. Conversely, arthropod classification measures in-domain performance where the underlying embedding space aligns with the pretraining distribution.

5.1 Fungi Classification

We use the MycoAI Benchmark dataset for model training and reference taxonomy construction. All reported results enforce self-match exclusion as detailed in Section 4.2.

Accuracy. Table 1 summarizes classification performance across MycoAI datasets. Barcode foundation models generalize effectively to fungal sequences, with BarcodeMamba outperforming tra-

Table 1: Performance of PROTAX-SSL variants and baselines on three datasets for fungi classification across taxonomic levels (family, genus, species). PROTAX models are reported as mean accuracy ± 1 standard deviation across 5 random seeds. **Bold**: best; underlined: second best.

Model	Accuracy (%)								
	MycoAI Benchmark			Yeast			Filamentous		
	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.
BLAST	94.7	93.1	55.0	86.6	92.9	75.4	81.4	71.5	33.4
MycoAI-CNN	93.9	87.8	57.1	90.5	86.4	60.0	84.1	69.8	28.2
MycoAI-BERT	93.2	80.3	39.3	88.9	75.7	33.5	85.1	60.8	16.6
CNN Encoder	97.5	93.6	72.6	94.1	88.3	67.6	84.5	69.1	31.4
BarcodeBERT (fine-tuned)	97.8	92.0	58.9	95.4	88.6	59.1	87.8	70.2	27.7
BarcodeBERT (k-NN)	87.7	83.3	60.9	82.9	79.6	62.9	80.6	69.2	36.7
BarcodeMamba (k-NN)	98.2	95.8	77.3	99.1	98.6	86.4	95.9	89.2	57.4
PROTAX-GPU	85.0± 1.5	79.3± 0.6	55.0± 0.68	72.0± 0.08	64.1± 0.05	46.4± 0.88	75.0± 0.59	60.0± 0.92	27.0± 0.47
PROTAX[BarcodeBERT]	87.8± 0.01	83.4± 0.01	61.4± 0.02	83.1± 0.04	79.6± 0.04	63.3± 0.03	80.7± 0.02	69.2± 0.02	37.0± 0.04
PROTAX[BarcodeMamba]	98.3± 0.01	96.0± 0.01	78.1± 0.04	**99.3± 0.01**	**98.7± 0.02**	**87.6± 0.07**	96.0± 0.03	**89.3± 0.05**	57.4± 0.14
PROTAX[BarcodeBERT+Mamba]	98.4± 0.01	96.1± 0.01	79.0± 0.03	99.2± 0.02	98.5± 0.07	86.2± 0.1	**96.2± 0.09**	89.2± 0.15	57.5± 0.26
PROTAX[BarcodeBERT+Mamba, MLP]	**98.5± 0.01**	**96.3± 0.01**	**80.4± 0.02**	**99.3± 0.00**	98.6± 0.01	85.7± 0.04	**96.2± 0.03**	89.2± 0.07	**58.1± 0.11**

ditional and supervised baselines. Self-supervised representations substantially improve PROTAX classification performance: all PROTAX-SSL variants outperform PROTAX-GPU, with the MLP variant achieving a 25.4 percentage points (p.p.) gain on the MycoAI Benchmark at the species-level. Integrating self-supervised representations into PROTAX preserves or slightly improves accuracy relative to the underlying encoders. Combining BarcodeBERT and BarcodeMamba spaces further demonstrates complementary information: the MLP variant mostly achieves the best results and improves accuracy by 3.1 p.p. over BarcodeMamba (k-NN) on the MycoAI Benchmark at the species-level. Conversely, the Yeast set forms a concentrated fungal clade with simpler decision boundaries, favoring the single-encoder `PROTAX[BarcodeMamba]`.

To assess generalization of results on MycoAI Benchmark, we evaluated an 80/20 species-stratified split set-up (Appendix D). All PROTAX-SSL variants showed nearly identical training and test accuracies. We therefore retained evaluation on the full dataset for better comprehension.

Predictive Reliability. Figure 2 demonstrates the calibration advantages of PROTAX-SSL. Encoders with k-NN exhibit pronounced overconfidence, most evident at the species level. PROTAX-SSL variants mitigate the bias of their corresponding encoders, significantly improving probability calibration across ranks. Furthermore, multi-embedding variants achieve the best calibration, indicating that complementary representations not only boost accuracy but also refine predictive uncertainty estimation, with the MLP model yielding near-ideal calibration in most cases.

5.2 Unknown Species

Since PROTAX explicitly models "unknown" taxa, it provides a natural baseline for both detection and placement of unseen species. While BarcodeBERT and BarcodeMamba conducted k-NN-based placement experiments on unseen species, they do not explicitly model unknown taxa and therefore do not have a natural way to perform detection. Table 2 summarizes the performance on unknown

Table 2: Performance of PROTAX-SSL variants and PROTAX-GPU on detecting *unknown* species and placing them under the correct parent genus, reported as mean accuracy ± 1 standard deviation across 5 random seeds. **Bold**: best; underlined: second best.

Model	Accuracy (%)	
	Detection	Placement
PROTAX-GPU	85.5± 2.6	36.0± 2.5
PROTAX[BarcodeBERT]	95.5± 0.22	45.9± 1.0
PROTAX[BarcodeMamba]	97.3± 1.2	76.6± 0.89
PROTAX[BarcodeBERT+Mamba]	**99.9± 0.02**	80.2± 0.1
PROTAX[BarcodeBERT+Mamba, MLP]	**99.9± 0.02**	**80.6± 0.05**

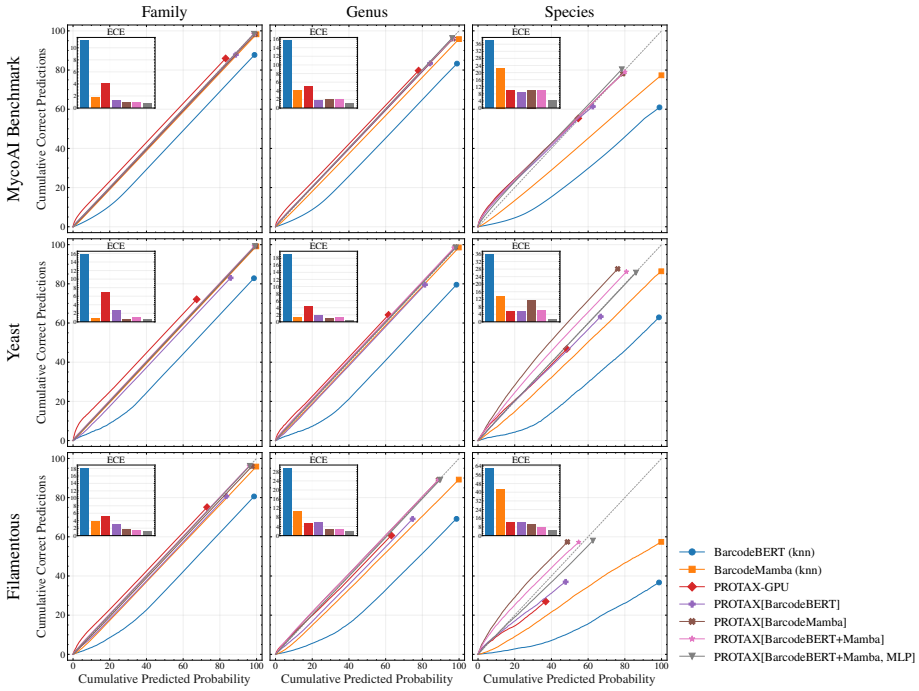

Figure 2: Calibration of PROTAX-SSL variants and PROTAX-GPU on the MycoAI datasets. Rows show different datasets, while columns show taxonomic levels. A perfectly calibrated model follows the diagonal. Overconfident and underconfident models lie below and above the diagonal, respectively. The top-left inset on each plot shows expected calibration error (ECE), where lower values indicate better calibration. Best viewed in colour.

species. Consistent with earlier classification results, all PROTAX-SSL variants substantially outperform PROTAX-GPU; notably, `PROTAX[BarcodeMamba]` increases detection by 11.8 p.p. and more than doubles placement accuracy. Multi-encoder variants achieve near-perfect detection, while `PROTAX[BarcodeBERT+Mamba, MLP]` attains the highest placement accuracy overall.

Accurate genus-level placement of unknown species demonstrates taxonomic clusters that generalize beyond seen species. PROTAX-SSL not only leverages these clusters provided by SSL representations to surpass PROTAX-GPU in open-set recognition, but also exploits multiple embedding spaces to improve generalization to unseen species beyond what each individual space can achieve.

5.3 Arthropod Classification

Table 3 summarizes the classification performance on the arthropod datasets. On the Canadian Invertebrates, we observe trends similar to those on MycoAI. However, because the stand-alone encoders already achieve strong performance, there is less room for further improvement with PROTAX-SSL. Multi-embedding variants provide modest gains at the species level.

A contrasting trend emerges on FinBOL, where PROTAX-GPU attains the highest family- and genus-level accuracies and remains competitive at the species level. This indicates that in tightly curated reference libraries with minimal label noise, raw sequence space provides sufficient discriminative signals, leaving limited room for learned representations to add value. Embedding projections may introduce subtle approximation artifacts without adding useful information (see Appendix F for analysis of taxonomic clusters quality of raw and embedding spaces). Nevertheless, `PROTAX[BarcodeBERT]` achieves the highest species-level accuracy, demonstrating that learned representations retain an edge for fine-grained discrimination among closely related taxa.

Unlike on MycoAI and Canadian Invertebrates, the multi-embedding variants on FinBOL provide no advantage over single-encoder models. Similar to results on Yeast, it suggests that in simpler classification settings, a single well-aligned representation space can capture the relevant taxonomic infor-

Table 3: Performance of PROTAX-SSL variants and baselines on two datasets for arthropod classification across taxonomic levels (family, genus, species). **Bold**: best; underlined: second best.

Model	Accuracy (%)								
	Can. Inv. Train			Can. Inv. Test			FinBOL		
	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.
BarcodeBERT (k-NN)	99.8	99.6	98.5	99.8	99.6	98.5	97.9	95.8	85.2
BarcodeMamba (k-NN)	99.6	99.3	98.3	99.6	99.3	98.3	96.5	94.6	84.7
PROTAX-GPU	95.0	90.7	84.3	94.9	90.8	84.4	**98.4**	**96.4**	86.5
PROTAX[BarcodeBERT]	99.8	99.6	98.5	99.8	99.6	98.5	97.9	95.8	**86.9**
PROTAX[BarcodeMamba]	99.6	99.3	98.3	99.6	99.3	98.3	96.5	94.5	85.6
PROTAX[BarcodeBERT+Mamba]	**99.9**	99.7	98.9	**99.9**	99.7	98.7	98.1	95.0	86.5
PROTAX[BarcodeBERT+Mamba, MLP]	**99.9**	**99.8**	**99.2**	**99.9**	**99.8**	**99.2**	98.2	96.0	86.3

mation, while concatenating additional embeddings may introduce redundancy or noise that offsets the potential benefits of combining representations. This suggests that the effectiveness of combining multiple embedding spaces depends on the complexity of the underlying classification task. Finally, predictive reliability on arthropods aligns with earlier findings on MycoAI (Appendix E).

5.4 Discussion and Limitations

We view PROTAX-SSL as a framework for enhancing foundation models' reliability and predictive capabilities, beyond what typical downstream classifiers can provide. Across all benchmarks, our framework delivers critical advantages: it provides significant calibration correction and a robust mechanism for open-set recognition, without sacrificing classification accuracy. It also incorporates information about the underlying taxonomy and produces predictions grounded in its hierarchy.

The performance of the multi-embedding MLP further highlights the framework's ability to leverage heterogeneous features to achieve gains beyond those attainable from individual representations alone. This modularity opens opportunities to explore broader combinations of features, particularly multimodal inputs that integrate representations of specimen images and DNA barcodes.

A limitation of PROTAX-SSL is its dependence on the underlying encoders. When self-supervised representations provide little to no advantage over raw sequence space, PROTAX-GPU may offer a sufficient alternative. The embeddings advantage emerges most clearly at the species level and on larger, noisier datasets, the regime that dominates real-world biodiversity. Moreover, the flexibility to combine different features introduces a large space of configurations that must be evaluated empirically, making predictor selection a hyperparameter to consider when optimizing performance.

6 Conclusion

Reliable taxonomic classification requires more than accuracy. Predictions inform downstream ecological monitoring and policy, and the field's default reliance on linear probes and nearest-neighbour classifiers has left calibration and unknown-taxa handling largely unexamined. PROTAX-SSL addresses both: by coupling self-supervised DNA barcode representations with hierarchical probabilistic inference, it produces predictions whose probabilities track empirical correctness and whose probability mass can be allocated to taxa absent from the reference database.

By providing a framework for more reliable taxonomic classification, PROTAX-SSL can support biodiversity monitoring at the scale required by modern programs while helping distinguish confident identifications from uncertain cases, and advance novel species discovery. At the same time, its modular design allows researchers to adapt the framework to future advances in foundation models and extend it with new representations and sources of biological information, providing a foundation for continued research in reliable biodiversity analysis.

AI use statement

In this work, we used generative AI tools, mainly ChatGPT and Google Gemini, to assist with polishing the manuscript. First drafts were always written by the authors and then provided to these

models to correct grammatical errors, improve flow, and enhance clarity. We also used these tools to assist with the discovery and retrieval of related work. Any sources suggested by the models were deliberately reviewed and validated by the authors before being included in the paper. Additionally, we used Cursor Agent to assist with research execution, specifically for coding and debugging parts of the experiments and training pipeline. The tools were not used to implement any component from scratch. All AI-assisted code edits were thoroughly reviewed and revised by the authors before acceptance and use.

We did not use generative AI tools to formulate or prove mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret results. Generating synthetic datasets and translating content are not applicable to this work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

All datasets used in this work are publicly available from their respective original sources. We intend to make the complete codebase for embedding generation, reference taxonomy construction, training of all model variants, and evaluation publicly available. Training configurations and computational resources are provided in Appendix B, while details of the MLP architecture are provided in Appendix C.

References

Kessy Abarenkov, Allan Zirk, Timo Piirmann, Raivo Pöhönen, Filipp Ivanov, R. Henrik Nilsson, and Urmas Kõljalg. Full unite+insd dataset for fungi, 2023. URL https://doi.org/10.15156/BIO/2938065.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2.

Frederic Austerlitz, Olivier David, Brigitte Schaeffer, Kevin Bleakley, Madalina Olteanu, Raphael Leblois, Michel Veuille, and Catherine Laredo. DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics*, 10(S14):S10, 11 2009. doi: 10.1186/1471-2105-10-s14-s10. URL https://doi.org/10.1186/1471-2105-10-s14-s10.

Sarkhan Badirli, Zeynep Akata, George Mohler, Christine Picard, and Mehmet M. Dundar. Fine-grained zero-shot learning with dna as side information. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy S. Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 19352–19362. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/a18630ab1c3b9f14454cf70dc7114834-Paper.pdf.

Yves Basset, Lukas Cizek, Philippe Cuénoud, Raphael K. Didham, François Guilhaumon, Olivier Missa, Vojtech Novotny, Frode Ødegaard, Tomas Roslin, Jürgen Schmidl, Alexey K. Tishechkin, Neville N. Winchester, David W. Roubik, Henri-Pierre Aberlenc, Johannes Bail, Héctor Barrios, Jon R. Bridle, Gabriela Castaño-Meneses, Bruno Corbara, Gianfranco Curletti, Wesley Duarte da Rocha, Domir De Bakker, Jacques H. C. Delabie, Alain Dejean, Laura L. Fagan, Andreas Floren, Roger L. Kitching, Enrique Medianero, Scott E. Miller, Evandro Gama de Oliveira, Jérôme Orivel, Marc Pollet, Mathieu Rapp, Sérvio P. Ribeiro, Yves Roisin, Jesper B. Schmidt, Line Sørensen, and Maurice Leponce. Arthropod diversity in a tropical forest. *Science*, 338(6113): 1481–1484, 2012. doi: 10.1126/science.1226727. URL https://www.science.org/doi/abs/10.1126/science.1226727.

Adam L. Bazinet and Michael P. Cummings. A comparative evaluation of sequence classification programs. *BMC Bioinformatics*, 13:92, May 2012. doi: 10.1186/1471-2105-13-92.

David Bickford, David J. Lohman, Navjot S. Sodhi, Peter K. L. Ng, Rudolf Meier, Kevin Winker, Krista K. Ingram, and Indraneil Das. Cryptic species as a window on diversity and conservation.

Trends in Ecology & Evolution, 22(3):148–155, 2007. ISSN 0169-5347. doi: 10.1016/j.tree.2006.11.004.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality, 2024. URL https://arxiv.org/abs/2405.21060.

Jeremy R. deWaard, Sujeevan Ratnasingham, Evgeny V. Zakharov, Alex V. Borisenko, Dirk Steinke, Angela C. Telfer, Kate H. J. Perez, Jayme E. Sones, Monica R. Young, and Valerie Levesque-Beaudin. A reference library for canadian invertebrates with 1.5 million barcodes, voucher specimens, and dna samples. *Scientific Data*, 6(1):308, 2019. doi: 10.1038/s41597-019-0320-2. URL https://doi.org/10.1038/s41597-019-0320-2.

Andrew Dopheide, Leah K. Tooman, Stefanie Grosser, Barbara Agabiti, Birgit Rhode, Dong Xie, Mark I. Stevens, Nicola Nelson, Thomas R. Buckley, Alexei J. Drummond, and Richard D. Newcomb. Estimating the biodiversity of terrestrial invertebrates on a forested island using dna barcodes and metabarcoding data. *Ecological Applications*, 29(4):e01877, 2019. doi: 10.1002/eap.1877. URL https://doi.org/10.1002/eap.1877.

Tiancheng Gao and Graham W. Taylor. Barcodemamba: State space models for biodiversity analysis, 2024. URL https://arxiv.org/abs/2412.11084.

Tiancheng Gao, Scott C. Lowe, Brendan Furneaux, Angel X Chang, and Graham W. Taylor. Barcodemamba+: Advancing state-space models for fungal biodiversity research, 2025. URL https://arxiv.org/abs/2512.15931.

Ajay Kumar Gautam, Rajnish Kumar Verma, Shubhi Avasthi, Sushma, Yogita Bohra, Bandarupalli Devadatha, Mekala Niranjan, and Nakarin Suwannarach. Current Insight into Traditional and Modern Methods in Fungal Diversity Estimates. *J. Fungi*, 8:226–, 2022. URL https://doi.org/10.3390/jof8030226.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Paul D. N. Hebert, Alina Cywinska, Shelley L. Ball, and Jeremy R. deWaard. Biological identifications through dna barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270 (1512):313–321, 02 2003. doi: 10.1098/rspb.2002.2218. URL https://doi.org/10.1098/rspb.2002.2218.

P. M. Hollingsworth, A. A. Clark, L. L. Forrest, J. Richardson, R. T. Pennington, D. G. Long, R. Cowan, M. W. Chase, M. Gaudeul, and M. L. Hollingsworth. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. *Molecular Ecology Resources*, 9(2):439–457, March 2009. doi: 10.1111/j.1755-0998.2008.02439.x.

François Keck, Marjorie Couton, and Florian Altermatt. Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources*, 23(4): 742–755, 2023. doi: doi.org/10.1111/1755-0998.13746.

Roy Li, Sujeevan Ratnasingham, Iuliia Zarubiieva, Panu Somervuo, and Graham W. Taylor. Protaxgpu: a scalable probabilistic taxonomic classification system for dna barcodes. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379(1904):20230124, 05 2024. ISSN 0962-8436. doi: 10.1098/rstb.2023.0124. URL https://doi.org/10.1098/rstb.2023.0124.

Pablo Millan Arias, Niousha Sadjadi, Monireh Safari, ZeMing Gong, Austin T Wang, Joakim Bruslund Haurum, Iuliia Zarubiieva, Dirk Steinke, Lila Kari, Angel X Chang, Scott C Lowe, and Graham W Taylor. Barcodebert: transformers for biodiversity analyses. *Bioinformatics Advances*, 6 (1), 2026. ISSN 2635-0041. doi: 10.1093/bioadv/vbag054. URL http://dx.doi.org/10.1093/bioadv/vbag054.

Camilo Mora et al. How many species are there on earth and in the ocean? *PLoS Biology*, 9(8): e1001127, 2011.

Masatoshi Nei and Sudhir Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, New York, 2000. ISBN 9780195135848. doi: 10.1093/oso/9780195135848.001.0001.

S. L. Pimm, C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, and J. O. Sexton. The biodiversity of species and their rates of extinction, distribution, and protection. *Science*, 344(6187):1246752, 5 2014. doi: 10.1126/science.1246752. URL https://doi.org/10.1126/science.1246752.

Sujeevan Ratnasingham and Paul D. N. Hebert. bold: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, 7(3):355–364, 2007. doi: https://doi.org/10.1111/j.1471-8286.2007.01678.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1471-8286.2007.01678.x.

Luuk Romeijn, Andrius Bernatavicius, and Duong Vu. Mycoai: Fast and accurate taxonomic classification for fungal its sequences. *Molecular Ecology Resources*, 24(8):e14006, 2024. doi: 10.1111/1755-0998.14006.

Tomas Roslin, Panu Somervuo, Mikko Pentinsaari, Paul D. N. Hebert, Jireh Agda, Petri Ahlroth, Perttu Anttonen, Jouni Aspi, Gergin Blagoev, et al. A molecular-based identification resource for the arthropods of finland. *Molecular Ecology Resources*, 22(2):803–822, 2022. doi: 10.1111/1755-0998.13510.

Howard A. Ross, Sumathi Murugan, and Wai Lok Sibon Li. Testing the reliability of genetic methods of species identification via simulation. *Systematic Biology*, 57(2):216–230, 4 2008. doi: 10.1080/10635150802032990. URL https://pubmed.ncbi.nlm.nih.gov/18398767/.

Osvaldo E. Sala, F. Stuart Chapin, Iii, Juan J. Armesto, Eric Berlow, Janine Bloomfield, Rodolfo Dirzo, Elisabeth Huber-Sanwald, Laura F. Huenneke, Robert B. Jackson, Ann Kinzig, Rik Leemans, David M. Lodge, Harold A. Mooney, Martín Oesterheld, N. LeRoy Poff, Martin T. Sykes, Brian H. Walker, Marilyn Walker, and Diana H. Wall. Global Biodiversity Scenarios for the year 2100. *Science*, 287(5459):1770–1774, 3 2000. doi: 10.1126/science.287.5459.1770. URL https://doi.org/10.1126/science.287.5459.1770.

C. L. Schoch, K. A. Seifert, S. Huhndorf, V. Robert, J. L. Spouge, C. A. Levesque, W. Chen, Fungal Barcoding Consortium, and Fungal Barcoding Consortium Author List. Nuclear ribosomal internal transcribed spacer (its) region as a universal dna barcode marker for fungi. *Proceedings of the National Academy of Sciences of the United States of America*, 109(16):6241–6246, April 2012. doi: 10.1073/pnas.1117018109.

Geethen Singh, Glenn Moncrieff, Zander Venter, Kerry Cawse-Nicholson, Jasper Slingsby, and Tamara B. Robinson. Uncertainty quantification for probabilistic machine learning in earth observation using conformal prediction. *Scientific Reports*, 14(1):16166, 7 2024. doi: 10.1038/s41598-024-65954-w. URL https://www.nature.com/articles/s41598-024-65954-w.

Panu Somervuo, Satu Koskela, Jussi Pennanen, R. Henrik Nilsson, and Otso Ovaskainen. Unbiased probabilistic taxonomic classification for dna barcoding. *Bioinformatics*, 32(19):2920–2927, October 2016. doi: 10.1093/bioinformatics/btw346.

Panu Somervuo, Douglas W. Yu, Charles C. Y. Xu, Yinqiu Ji, Jenni Hultman, Helena Wirta, and Otso Ovaskainen. Quantifying uncertainty of taxonomic placement in dna barcoding and metabarcoding. *Methods in Ecology and Evolution*, 8(4):398–407, 2017. doi: 10.1111/2041-210x.12721. URL https://doi.org/10.1111/2041-210x.12721.

Duc Vu, Marizeth Groenewald, Szabolcs Szöke, Giuseppe Cardinali, Ursula Eberhardt, J. B. Stielow, Martijn de Vries, G. J. M. Verkleij, Pedro W. Crous, Teun Boekhout, and Vincent Robert. Dna barcoding analysis of more than 9,000 yeast isolates contributes to quantitative thresholds for yeast species and genera delimitation. *Studies in Mycology*, 85:91–105, 2016. doi: 10.1016/j.simyco.2016.11.007. URL https://doi.org/10.1016/j.simyco.2016.11.007.

Duc Vu, Marizeth Groenewald, Martijn de Vries, Tim Gehrmann, J. B. Stielow, Ursula Eberhardt, Abdallah M. S. Al-Hatmi, Johannes Z. Groenewald, Giuseppe Cardinali, Jos Houbraken, Teun

Boekhout, Pedro W. Crous, Vincent Robert, and Gerard J. M. Verkley. Large-scale generation and analysis of filamentous fungal dna barcodes boosts coverage for kingdom fungi and reveals thresholds for fungal species and higher taxon delimitation. *Studies in Mycology*, 92: 135–154, 2019. doi: 10.1016/j.simyco.2018.05.001. URL https://doi.org/10.1016/j.simyco.2018.05.001.

Duong Vu, R. Henrik Nilsson, and Gerard J. M. Verkley. Dnabarcoder: An open-source software package for analysing and predicting dna sequence similarity cutoffs for fungal sequence identification. *Molecular Ecology Resources*, 22(7):2793–2809, 2022. doi: 10.1111/1755-0998.13651.

Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naive bayesian classifier for rapid assignment of rrna sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, 73(16):5261–5267, 2007. doi: 10.1128/AEM.00062-07.

Shuaihua Ye, Katherine J. Siddle, Daniel J. Park, and Pardis C. Sabeti. Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794, August 2019. doi: 10.1016/j.cell.2019.07.010.

A Dataset Completeness and Overlaps

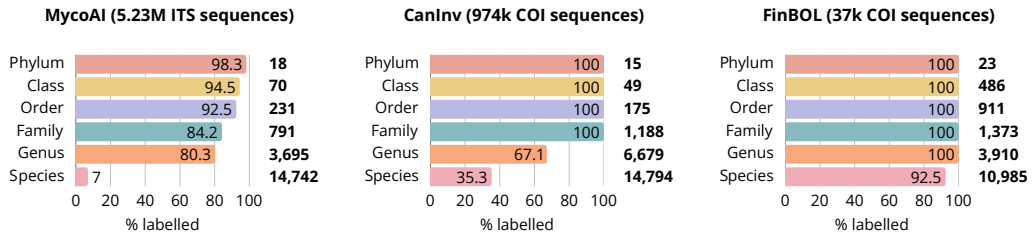

Figure 3: Overview of the datasets characteristics. It shows barcode type and scale, the number of represented taxa at each taxonomic level, and the proportion of sequences with available annotations. Best viewed in color.

Figure 3 summarizes the characteristics of the three mentioned datasets across taxonomic ranks: MycoAI contains broad fungal diversity but becomes increasingly incomplete at finer taxonomic ranks, particularly at the species level; Canadian Invertebrates provides complete higher-rank annotations but only partial genus- and species-level labels; and FinBOL is substantially smaller yet highly curated, with nearly complete annotations throughout the taxonomy.

Table 4: Species-level and identical-barcode overlap between the MycoAI Benchmark set used for training and the fungal test sets. Species overlap measures the proportion of species present in both the training and test sets, while identical-barcode overlap measures the proportion of test sequences whose exact barcode sequence is also present in the training set.

	Species Overlap			Identical Barcode Overlap		
Dataset	Total	Overlap	Overlap (%)	Total	Overlap	Overlap (%)
Validation	533	0.0	0.0	10,482	0.0	0.0
Yeast	1,157	616	53.24	4,235	3,673	86.73
Filamentous	5,537	2,650	47.86	10,721	695	6.48

Table 4 summarizes the species-level and identical-barcode overlap between the training set (MycoAI Benchmark) and the fungi classification test sets used in our work. The Yeast dataset occupies an intermediate position, exhibiting moderate species overlap (53.24%) but high barcode overlap (86.73%), assessing performance within a taxonomically concentrated fungal clade. In contrast, the Filamentous Fungi dataset contains 47.86% species overlap and only 6.48% identical-barcode overlap with the training set, making it the most challenging test set for fungi classification.

B Training Configurations and Computational Resources

Table 5: Summary of the hyperparameters and training times used for PROTAX-SSL variants and PROTAX-GPU. Batch sizes and training times differ across datasets due to GPU memory constraints and dataset size.

				Batch Size			Total Training Time (min)		
Model Variant	Base LR	Epochs	σ_β	MycoAI Benchmark	Canadian Inv.	FinBOL	MycoAI Benchmark	Canadian Inv.	FinBOL
PROTAX-GPU	5.0	20	5.0	512	2048	2048	105	10	7
PROTAX[BarcodeBERT]	5.0	20	5.0	512	2048	2048	102	9	7
PROTAX[BarcodeMamba]	5.0	20	5.0	512	2048	2048	102	9	7
PROTAX[BarcodeBERT+Mamba]	5.0	20	5.0	128	1024	1024	260	46	35
PROTAX-SSL[BarcodeBERT+Mamba, MLP]	0.005	40	0.05	64	64	64	572	105	85

Table 5 summarizes the learning parameters and time required to train PROTAX-SSL variants and PROTAX-GPU. All models were trained using stochastic gradient descent (SGD) with cross-entropy, using a linear warmup followed by a linear decay learning-rate schedule. Formally, the learning rate at training step t is defined as:

$$\eta(t) = \begin{cases} \eta_0 \dfrac{t}{T_w}, & t < T_w, \\ \eta_0 \left(1 - \dfrac{t - T_w}{T - T_w}\right), & T_w \leq t \leq T, \end{cases}$$

where η_0 denotes the base learning rate, T is the total number of training steps, and T_w is the number of warmup steps. In all experiments, the warmup period was set to 10% of the total training duration.

Hyperparameter tuning was performed using the MycoAI validation set. Resulting configuration was found to generalize well to the Canadian Invertebrates and FinBOL datasets. Therefore, to ensure consistency and avoid dataset-specific tuning, the same training parameters and initialization strategy were used across all datasets. However, because the linear and MLP variants represent fundamentally different model classes, architecture-dependent hyperparameters were tuned separately for each. For each classifier architecture, training was terminated after a fixed number of epochs that was empirically found to provide the best trade-off between accuracy and calibration. The selected epoch counts were used consistently across all experiments for the corresponding architecture.

All experiments were conducted on a SLURM HPC using a NVIDIA L40S GPU equipped with 256 GB of system memory. The implementation was developed in Python using JAX for model training and GPU acceleration. Batch size was selected as the largest value that could fit within the memory constraints of the available GPUs.

C Multilayer Perceptron Architecture

For the MLP variant, we employ a fully connected feed-forward neural network consisting of five hidden layers, each containing 256 neurons followed by ReLU activation functions. The chosen architecture represents a compromise between predictive performance, calibration quality, and computational efficiency. Architecture was selected with the objective of evaluating the effectiveness of integrating self-supervised DNA representations within the PROTAX framework, rather than exhaustively optimizing the neural network itself. The design space of alternative architectures, activation functions, and residual connections was not systematically explored. Investigating more specialized architectures that better exploit the structure of the predictor space may yield further improvements in both classification accuracy and probability calibration, and therefore represents an interesting direction for future research.

D MycoAI Benchmark Train-Test Split

To assess potential overfitting on the MycoAI Benchmark dataset, we conducted an additional experiment using an 80/20 species-stratified train-test split. The split preserves the species distribution across both subsets while ensuring that test sequences are excluded from the reference library during

evaluation. Reference taxonomy construction and model training were performed exclusively using the 80% training subset, while the remaining 20% was held out for evaluation. Training performance was evaluated using self-match exclusion.

Table 6: Performance of PROTAX-SSL variants and baselines on the MycoAI Benchmark Train and Test splits across taxonomic levels (family, genus, species). **Bold**: best; underlined: second best.

Model	MycoAI Benchmark Accuracy (%)					
	Train			Test		
	Fam.	Gen.	Sp.	Fam.	Gen.	Sp.
BarcodeBERT (knn)	86.5	81.7	58.6	86.5	81.9	59.3
BarcodeMamba (knn)	98.1	95.5	76.2	98.2	95.6	76.9
PROTAX-GPU	84.4	77.4	52.5	84.5	78.0	53.3
PROTAX[BarcodeBERT]	86.6	81.8	58.9	86.6	82.1	59.7
PROTAX[BarcodeMamba]	98.2	95.6	76.8	98.3	95.8	77.6
PROTAX[BarcodeBERT+Mamba]	**98.3**	95.8	77.8	98.3	95.9	78.4
PROTAX[BarcodeBERT+Mamba, MLP]	**98.3**	**96.0**	**78.5**	**98.4**	**96.1**	**79.2**

Table 6 reports classification accuracy on both splits. The results remain consistent between the training and held-out test sets, while mirroring the trends observed in the main evaluation: PROTAX-SSL yields comparable or modest gains over the corresponding standalone k-NN baselines, while multi-embedding variants achieve the highest accuracy.

Notably, test accuracy exceeds training accuracy across variants, mostly at the species level. This likely reflects sample variation from the stratified split, where the held-out partition contains more separable sequences that are easier to classify. Overall, the consistent trends and absence of accuracy gaps indicate no signs of overfitting, confirming that evaluation via self-match exclusion serves as an effective proxy for generalization beyond observed sequences. We therefore retain the full MycoAI Benchmark evaluation in the main experiments to provide a more comprehensive assessment.

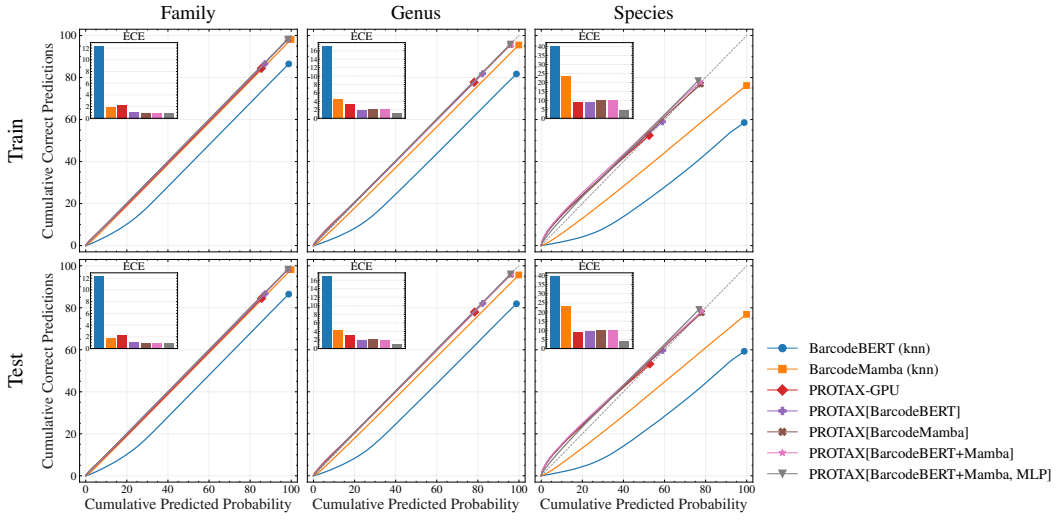

Figure 4: Calibration of PROTAX-SSL variants and PROTAX-GPU on MycoAI Benchmark Train and Test splits. Rows show different splits, while columns show performance at taxonomic levels. A perfectly calibrated model has a curve along the diagonal. Overconfident and underconfident models lie below and above the diagonal, respectively. Best viewed in colour.

Figure 4 compares the calibration of predictive probabilities. The curves remain closely aligned across the two splits, with no systematic shift toward better calibration on the training split. This provides evidence that our reported evaluation with self-match exclusion on the full MycoAI Bench-

mark does not produce artificially optimistic calibration estimates and that the observed calibration generalizes to sequences not seen during training.

E ARTHROPODS PREDICTIVE RELIABILITY

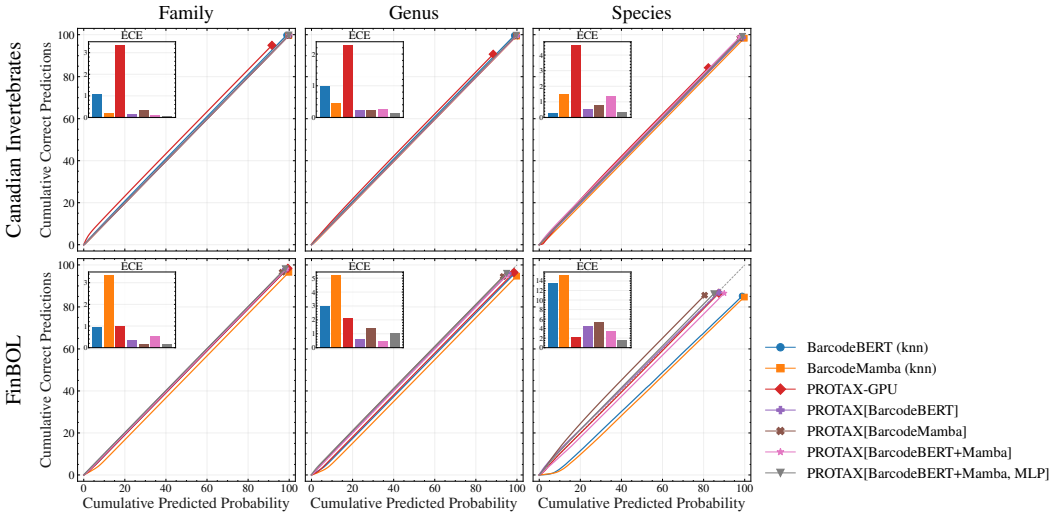

Figure 5: Calibration of PROTAX-SSL variants and PROTAX-GPU on the Arthropod datasets. Rows show different datasets, while columns show performance at taxonomic levels. A perfectly calibrated model has a curve along the diagonal. Overconfident and underconfident models lie below and above the diagonal, respectively. Best viewed in colour.

Figure 5 shows the calibration performance of the evaluated methods on the Canadian Invertebrates and FinBOL datasets. Unlike the fungal benchmarks, the differences between methods are relatively small. Since most models already achieve very high classification accuracies on both arthropod datasets, their calibration curves remain close to the ideal diagonal regardless of the specific classifier used. Nevertheless, PROTAX-SSL variants mostly correct calibration of their corresponding foundation models if needed, demonstrating that the calibration benefits of probabilistic inference still persist even when classification performance is near saturation. Overall, the arthropod results reinforce the findings from the fungal benchmarks: integrating self-supervised DNA barcode representations into the PROTAX framework produces more reliable uncertainty estimates.

F ANALYSIS OF REPRESENTATION SPACES

The results of PROTAX-GPU on FinBOL motivated a closer examination of how self-supervised representations and raw sequence space capture taxonomic structure. We compare cosine distances in the BarcodeMamba embedding space with p-distances in raw sequence space on both the FinBOL and MycoAI Benchmark datasets.

Figure 6 demonstrates the distributions of intra-species and inter-species distances between sequences. for the MycoAI dataset, BarcodeMamba provides a much better separation between the two distributions compared to the raw space, in which distances between sequences of different species totally overlap with sequences from the same species. On the other hand, for the FinBOL, raw space distances are slightly better separated than barcodemamba, explaining the better results PROTAX-GPU achieved.

Figure 7 further demonstrates how accuracy corresponds to the taxonomic clusters formed by the two spaces. Raw-sequences are not vectors and can't be represented with centroids, hence we represent their clusters with medoid radius of p-distances between the sequences inside each species. We observe same trend, where barcodemamba creates a much better space for mycoai sequences, and the raw space creates poor taxonomic clusters for which species clusters have big radiuses, which

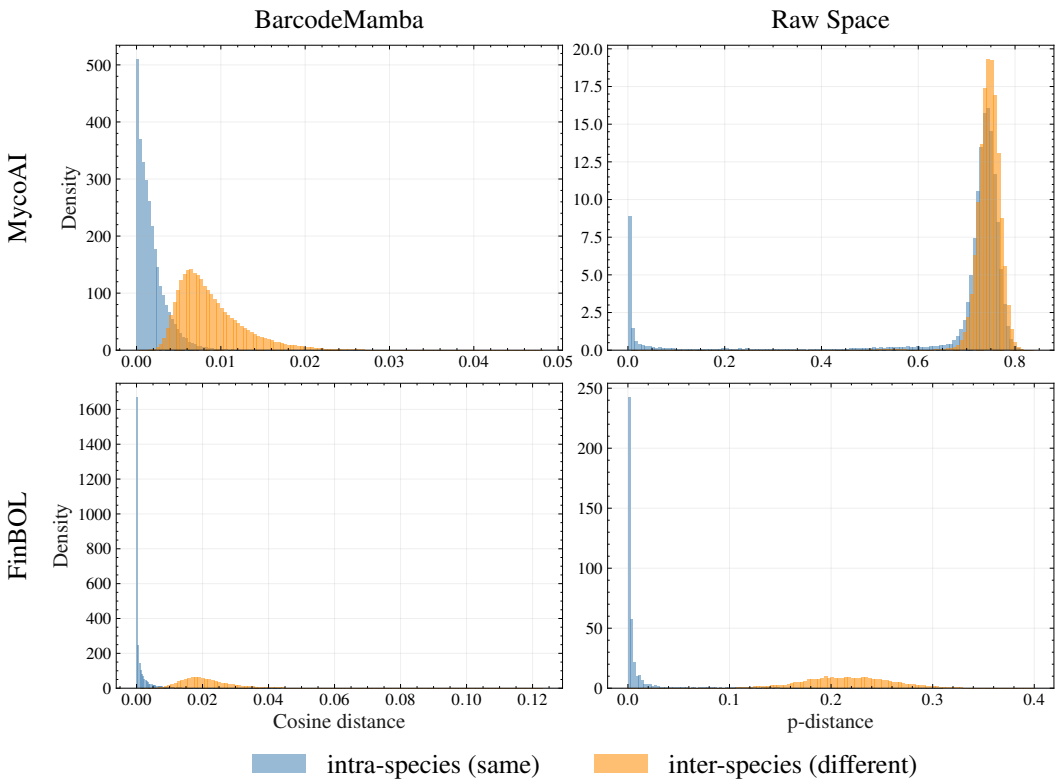

Figure 6: Distribution overlap between intra- and inter-species pairwise distances. Rows show datasets, and columns compare BarcodeMamba to the raw space. Best viewed in colour.

make the possiblity of them to overlap with each other high and hard to get classification borders that accurately separate them. On FinBOL, barcodemamba doesn't improve over the raw space.

These analysis highlight when self-supervised representations can become useful, which is when the data is complicated with high class imbalance and poor taxonomic clusters (which is the case in real-world taxonomy). DNA barcode encoders can project the sequences to another space that caputres taxonomic relationships much better and separate between species well. On the other hand, in the rare specific case of the raw space already providing a solid space of those clusters, SSL don't provide further improvement.

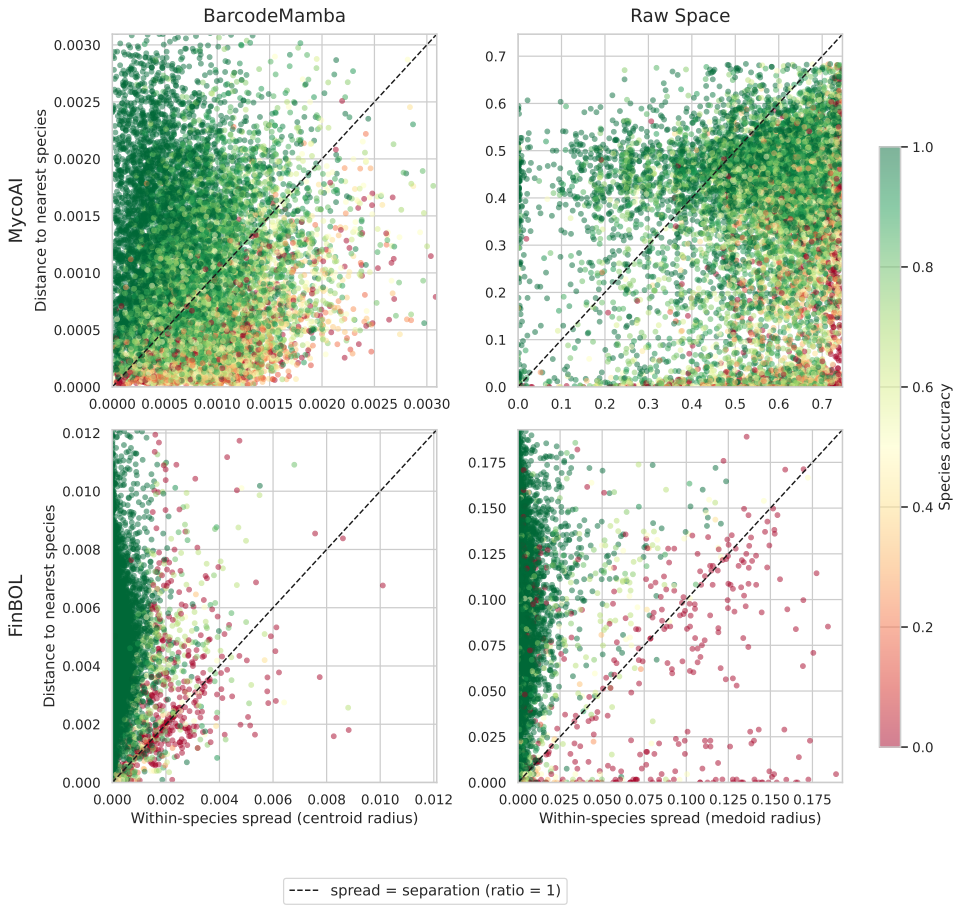

Figure 7: Within-species spread versus nearest-neighbour distance. Rows represent datasets, and columns compare BarcodeMamba to the raw space. Each point represents a species, coloured by classification accuracy. Best viewed in colour.